Robust Estimation of cDNA Microarray Intensities with Replicates

Raphael Gottardo, Adrian E. Raftery, Ka Yee Yeung and Roger E. Bumgarner [1]

Technical Report no. 438
Department of Statistics
University of Washington

December 1, 2003

[1]Raphael Gottardo is PhD Candidate, Department of Statistics, University of Washington, Box 354322 Seattle, WA 98195-4322 (E-mail: *raph@stat.washington.edu*; Web: *www.stat.washington.edu/raph*). Adrian E. Raftery is Professor of Statistics and Sociology, University of Washington, Department of Statistics Box 354322, Seattle, WA 98195-4322 (E-mail: *raftery@stat.washington.edu*; Web: *www.stat.washington.edu/raftery*). Ka Yee Yeung is Research Scientist, Department of Microbiology, University of Washington, Box 358070, Seattle, WA 98195. (E-mail: *kayee@u.washington.edu*). Roger Bumgarner is Assistant Professor, Department of Microbiology, University of Washington, Box 358070, Seattle, WA 98195. (E-mail: *rogerb@u.washington.edu*). The authors thank Julian Besag for helpful discussion and Angelique van't Wout for providing us with some of the data. This research was supported by NIH Grant 8 R01 EB002137-02, and Raftery's research was also partially supported by ONR Grant N00014-01-10745. Yeung and Bumgarner were supported by NIH-NIDDK grant 5U24DK058813-02.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 DEC 2003		**00-12-2003 to 00-12-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Robust Estimation of cDNA Microarray Intensities with Replicates	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Washington,Department of Statistics,Box 354322,Seattle,WA,98195-4322**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**27**	

Abstract

We consider robust estimation of gene intensities from cDNA microarray data with replicates. Several statistical methods for estimating gene intensities from microarrays have been proposed, but there has been little work on robust estimation of the intensities. This is particularly relevant for experiments with replicates, because even one outlying replicate can have a disastrous effect on the estimated intensity for the gene concerned. Because of the many steps involved in the experimental process from hybridization to image analysis, cDNA microarray data often contain outliers. For example, an outlying data value could occur because of scratches or dust on the surface, imperfections in the glass, or imperfections in the array production. We develop a Bayesian hierarchical model for robust estimation of cDNA microarray intensities. Outliers are modeled explicitly using a t-distribution, and our model also addresses classical issues such as design effects, normalization, transformation, and nonconstant variance. Parameter estimation is carried out using Markov Chain Monte Carlo.

The method is illustrated using two publicly available gene expression data sets. The between-replicate variability of the intensity estimates is reduced by 64% in one case and by 83% in the other compared to raw log ratios. The method is also compared to the ANOVA normalized log ratio, the removal of outliers based on Dixon's test, and the lowess normalized log ratio, and the between-replicate variation is reduced by more than 55% relative to the best of these methods for both data sets.

We also address the issue of whether the image background should be removed when estimating intensities. It has been argued that one should not do so because it increases variability, while the arguments for doing so are that there is a physical basis for the image background, and that not doing so will bias the estimated log-ratios of differentially expressed genes downwards. We show that the arguments on both sides of this debate are correct for our data, but that by using our model one can have the best of both worlds: one can subtract the background without greatly increasing variability.

KEY WORDS: Bayesian hierarchical model; Gene expression; Heteroscedasticity; Markov chain Monte Carlo; Outlier; t distribution.

Contents

List of Tables

List of Figures

1 INTRODUCTION

cDNA microarrays consist of thousands of individual DNA sequences printed in a high density array on a glass microscope slide using a robotic arrayer. A microarray works by exploiting the ability of a given labelled cDNA molecule to bind specifically to, or hybridize to, a complemetary sequence on the array. By using an array containing many DNA samples, scientists can measure—in a single experiment—the expression levels of hundreds or thousands of genes within a cell by measuring the amount of labelled cDNA bound to each site on the array. In a typical two-color microarray experiment, two mRNA samples, from control and treatment situations, are compared for gene expression. Treatment is taken in a broad sense to mean any condition different from the control. Both mRNA samples, or targets, are reverse-transcribed into cDNA, labeled using different fluorescent dyes (red and green dyes), then mixed and hybridized with the arrayed DNA sequences. The hybridized arrays are then imaged to measure the red and green intensities for each spot on the glass slide. Image analysis is an important aspect of microarray experiments, whose purpose is to provide estimates of the foreground and background intensities for both the red and green channels (Yang, Buckley, Dudoit, and Speed 2002). The estimates of the red and green intensities are the starting point of any statistical analysis such as testing for differential expression, discriminant analysis, and clustering (Yeung, Fraley, Murua, Raftery, and Ruzzo 2001; Gottardo, Pannucci, Kuske, and Brettin 2003).

In order to measure gene expression changes accurately, it is important to take into account the random and systematic variations that occur in every microarray experiment. One way to measure the variation is to use replicated experiments in which each gene is replicated several times. In recent years, there has been a considerable amount of work on the estimation of gene intensities and the detection of differentially expressed genes (Chen, Dougherty, and Bittner 1997; Newton, Kendziorski, Richmond, Blattner, and Tsui 2001; Dudoit, Yang, Callow, and Speed 2002). However, not much work has been done on robust estimation of the intensities. Because of the large number of steps involved in the experimental process from hybridization to image analysis, cDNA microarray data often contain outliers. For example, an outlying data value could occur because of scratches or dust on the surface, imperfections in the glass, imperfections in the array production, and so on. Outliers can also occur when the estimated difference between the background and the foreground intensities from the image analysis is small. There is a need for robust estimates of the intensities.

Some work has been done on quality measure and filtering. Such approaches consist of

calculating quality measures (sometime referred to as quality indixes) for a spot. Spots with low quality indices are usually removed. Among others, Tseng, Oh, Rohlin, Liao, and Wong (2001) filter genes based on the coefficient of variation, Lönnstedt and Speed (2002) remove spots with low intensities, and Ideker, Thorsson, Siegel, and Hood (2000) remove outliers using Dixon's test at the 10% level. However, in general, spots can fall anywhere in the range from "good to "bad" and the cut-offs used to judge the quality are somewhat arbitrary. Given that microarray experiments are expensive and that the number of replicates tends to be small, it would seem more satisfactory to downweight unreliable points than to discard them.

In this paper we introduce a Bayesian hierarchical model to estimate the intensities in a robust way. The robustness is achieved using a hierarchical-t formulation (Besag and Higdon 1999), which is more robust than the usual Gaussian model. Our model also deals with classical issues such as normalization, data transformation and nonconstant variance. The paper is organized as follows. Section 2 introduces the data structure and the notation. In Section 3 we present the Bayesian hierarchical model used to estimate the intensities and the parameter estimation method. In Section 4, we apply our model to experimental data and compare our results to those from popular alternative estimators. We also discuss whether one should subtract the background intensities from the image analysis. Finally, in Section 5 we discuss our results and possible extensions.

2 DATA

We used three data sets that are fairly typical of data in this area. All were produced by the University of Washington Center for Expression Arrays. They have the advantage that in each case we know whether or not all or some of the genes were differentially expressed.

The like and like data: This data set consists of 8 experiments using the same RNA preparation on 8 different slides. The samples that were applied to the arrays were RNA isolations from a HeLa cell line. The expression levels of about 7680 genes were measured. The same RNA was used for both samples. Theoretically, no genes should be differentially expressed.

The HIV1 data: This data set consists of four experiments using the same RNA preparation on 4 different slides. The expression levels of 4,608 cellular RNA transcripts were assessed in CD4-T-cell lines at time $t = 1$ hour after infection with HIV virus type 1. Included in this number was a set of 384 selected control genes spotted on each slide.

These included HIV-1 genes used as positive controls and nonhuman genes used as negative controls. Further details are given by van't Wout, Lehrma, Mikheeva, O'Keeffe, Katze, Bumgarner, Geiss, and Mullins (2003).

The HIV2 data: These data were collected in the same way and in the same laboratory as the HIV1 data, but using a different RNA preparation.

Each dataset is the result of a balanced dye-swap experiment. Two of the four replicates were hybridized with the green dye (Cy3) for the control and the red dye (Cy5) for the treatment; then the dyes were reversed on the other two replicates. The values in the Cy3 and Cy5 channels were extracted from each image using customized software written at the University of Washington (Spot-On Image, developed by R.E. Bumgarner and Erick Hammersmark). The image analysis provided four numbers for each gene in each replicate: a Green spot intensity, a Green background intensity, a Red spot intensity and a Red background intensity.

After the image analysis, the data take the form

$$(y_{iscr}, x_{iscr}),\ i = 1, \ldots, I;\ s = 1, 2;\ c = 1, 2;\ r = 1, \ldots, R$$

where y_{iscr} and x_{iscr} are the foreground and background intensities of gene i in the sample s with color c from the replicate r. Note that we use different indices for the color and the sample to allow dye-swap experiments.

Data transformation is an important initial step in microarray data analysis. It is often assumed that the log transformation of the raw data makes the effects additive. This assumption is approximately correct for gene expression data (Li and Wong 2000; Kerr, Martin, and Churchill 2000; Rocke and Durbin 2001; Dudoit, Yang, Callow, and Speed 2002). Figure 1(a) shows that on the log scale the dye effect is approximately additive. The dye effect is the result of an imbalance between the red and green intensities. It is known that such a dye bias can lead to a nonlinear effect (Dudoit, Yang, Callow, and Speed 2002). Figure 1(b) shows the log ratio intensity plotted against half the sum of the log intensities from the two channels; we will refer to the latter quantity as the "overall intensity." Figure 1(b) is just a 45° counterclockwise rotation of Figure 1(a). The Locally Weighted Scatterplot Smoother, or lowess (Cleveland 1979), indicates that such a nonlinear trend is present in the like and like data. We can see that the effects from the two different groups where the dyes have been swapped are almost identical but reversed. As a result, in a balanced dye-swap experiment, we expect the dye effect to be absent or at least greatly reduced when computing genewise

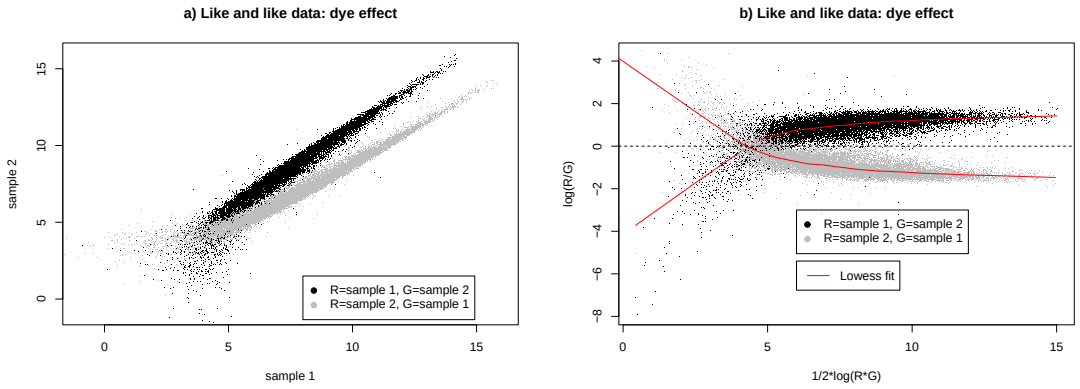

Figure 1: Effect of the Dye Swap on the Like and Like Data. (a) Log(sample 1) versus log(sample 2). This shows that the dye effect is approximately additive on the log scale. (b) Overall intensity versus log ratio. This shows that the non-linear effects will approximately cancel one another out in a balanced dye-swap experiment. The overall intensity is half the sum of the log intensities in the two channels, and all logarithms are to base 2.

averages. Throughout this paper, logarithms are to base 2, which is standard in the analysis of microarray data.

The log transformation usually stabilizes the variance for high intensity spots but low intensity spots can be highly variable. Figure 2 shows the sample standard deviations of all the genes as a function of their normalized log intensities. We normalized the intensities by subtracting (on the log scale) the design effects introduced in the next section, namely the sample, dye and replicate effects, and the dye-sample interaction effect.

The like and like data presented here measure the relative expression of a group of genes using the same mRNA in the two samples. As a result, we expect to observe the technical variation but not the biological variation. Figure 3 shows the normalized log intensities of ten different genes in two samples. Even though the data are normalized and there is no biological variation, we observe some outliers in each sample. It is clear that these outliers can have a big effect on the intensity estimates.

3 ROBUST INTENSITY ESTIMATION VIA BAYESIAN HIERARCHICAL MODELING

In this section we introduce the Bayesian hierarchical model we use to estimate the intensities in each sample. We use a Bayesian linear model (Lindley and Smith 1972) with t-distributed

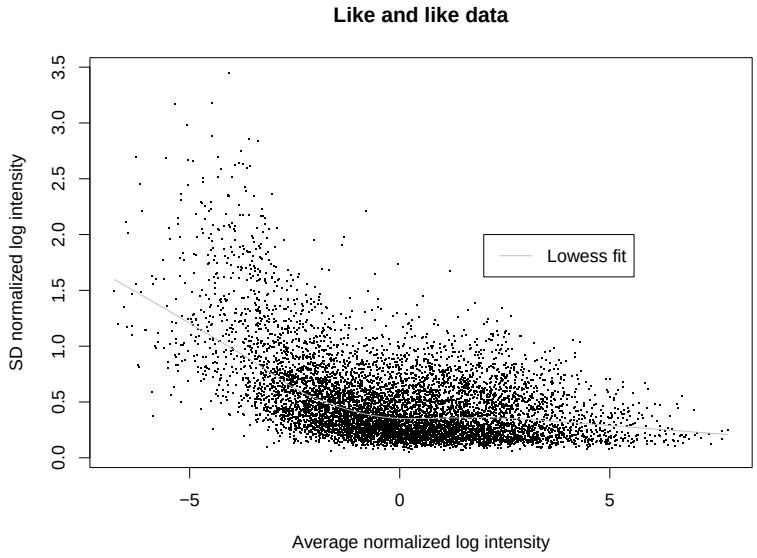

Figure 2: Sample Standard Deviation versus the Mean Intensity for Each Individual Gene of One Sample, after Normalization (i.e. after removing design effects). The variance is not constant and depends on the overall intensity.

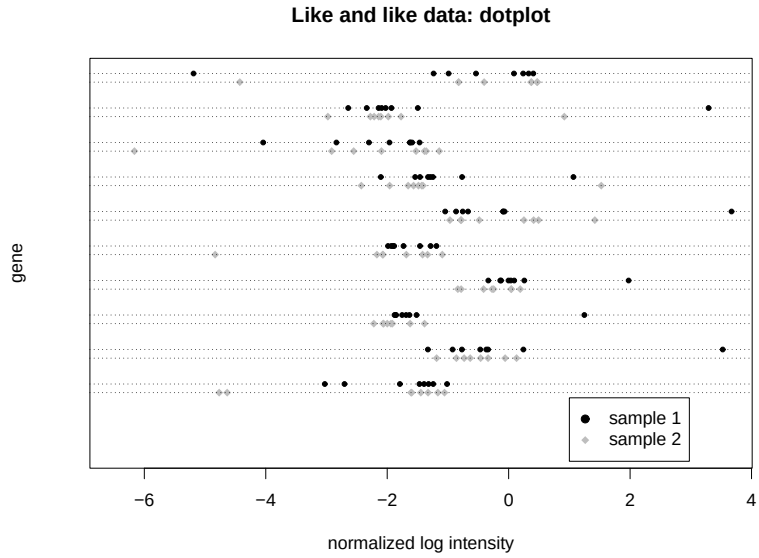

Figure 3: Dot Plots of 10 Genes from the Like and Like Data. Even though the data were normalized (i.e. the design effects were subtracted), some outliers are present in each sample.

sampling errors to allow for outliers (Besag and Higdon 1999). We also explicitly model the nonconstant variances by using an exchangeable prior for the gene precisions (Lewin, Richardson, Marshall, Glazier, and Aitman 2003). Our model includes design effects that deal with normalization issues (Kerr, Martin, and Churchill 2000). We model the intensities on the log scale because the effects are close to additive on that scale, as shown in Section 2, and because log measurements have a simple interpretation.

3.1 The Model

We model $y^*_{iscr} = \log_2(y_{iscr} - x_{iscr} + \kappa)$ where κ is a positive additive constant. This shifted logarithmic transformation was proposed by Tukey (1957) and studied in detail by Box and Cox (1964); it is often used to analyze gene expression data (Kerr, Martin, and Churchill 2000; Cui, Kerr, and Churchill 2002). Rocke and Durbin (2003) showed that the shifted logarithm can be an approximate variance-stabilizing transformation for gene expression data. The purpose of introducing the shift κ is to avoid taking the logarithm of negative numbers and to reduce the variance at low intensities. The parameter κ is estimated beforehand and is treated as fixed in the estimation of the full model, as described in Section 3.3.

Conditionally on the parameters $(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\delta}, \boldsymbol{\gamma})$, it is assumed that the $(y^*_{1scr}, y^*_{2scr})'$ are independent and can be written as

$$
\begin{aligned}
y^*_{iscr} = g_\kappa(y_{iscr} - x_{iscr}) &= \mu + \alpha_s + \beta_c + \eta_r + \gamma_{is} + \delta_{sc} + \frac{\epsilon_{iscr}}{\sqrt{w_{icr}}}, && (1)\\
(\gamma_{is}|\lambda_{\gamma_s}) &\sim \mathrm{N}(0, \lambda_{\gamma_s}), && (2)\\
(\epsilon_{i1cr}, \epsilon_{i2cr})'|\mathbf{V}_i &\sim \mathrm{N}_2(\mathbf{0}, \mathbf{V}_i), && (3)\\
(w_{icr}|\nu_r) &\sim \mathcal{G}a(\nu_r/2, \nu_r/2), && (4)
\end{aligned}
$$

where w_{icr} and $(\epsilon_{i1cr}, \epsilon_{i2cr})'$ are independent. Since the w's are independent of the ϵ's, we have $\frac{\epsilon_{iscr}}{\sqrt{w_{icr}}} \sim \mathcal{T}_{(\nu_r, \mathbf{0}, \mathbf{V}_i)}$, i.e. the (bivariate) errors have a bivariate t distribution with ν_r degrees of freedom and covariance matrix $\mathbf{V}_i$. The advantage of writing the model this way is that, conditioning on the w_{icr}, the sampling errors are again normal, but with different precisions. The interpretation is also easier, as, conditionally on the w_{icr}, estimation becomes a weighted least squares problem. The hierarchical structure of the model is summarized in the directed acyclic graph in Figure 4.

In (1), μ is the baseline intensity. The sample effect α_s is used to remove the bias between the two samples. If only a few of the genes are differentially expressed, the sample effect will measure only the sample bias and will not greatly affect the differentially expressed genes.

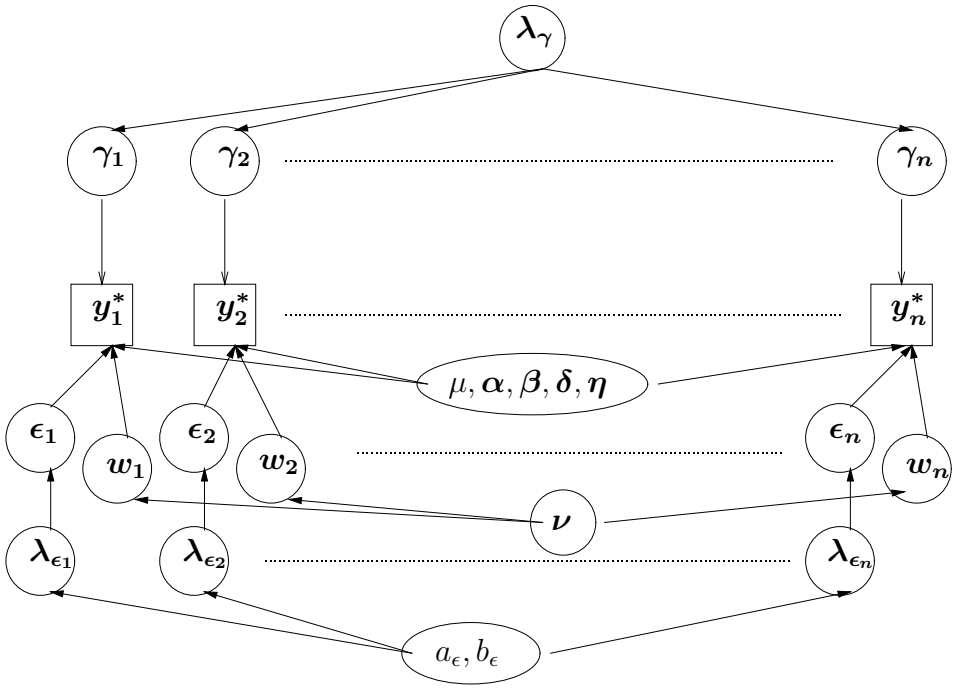

Figure 4: Directed Acyclic Graph of the General Model in Equation (1).

The dye effect is represented by β_c, and accounts for the well-known fact that the green dye (Cy3) tends to be brighter than the red dye (Cy5). The interaction of the sample s with the sample c is denoted by δ_{sc}, and is present because the different dyes tend to have different biases in different samples. The array effect of replicate r, η_r, is intended to normalize the overall intensity of each array across replicates. This parameter is needed because differences in overall intensity are frequent in microarray data. There are several reasons why this is so; for example, the amount of RNA solution used on each array might not be the same, leading to brighter arrays after the scanning process. Finally, γ_{is}, the effect of gene i in the sample s, is the quantity of interest. We model it as a random effect with a Gaussian distribution as defined by (2).

For a given gene, the correlation matrix, $\mathbf{V}_i$, allows the measurements from the two samples to be correlated. Ideker, Thorsson, Siegel, and Hood (2000) use a similar covariance structure in their linear model. The log transformation usually stabilizes the variance for high intensity spots but low intensity spots can be highly variable. Figure 2 shows that the variance is not constant across genes. A model that allows gene dependent variances seems more appropriate. The precision matrix (i.e. the inverse of the covariance matrix) is given

by

$$
\begin{aligned}
(\mathbf{V}_i^{-1}|\rho, \lambda_{\epsilon_{1i}}, \lambda_{\epsilon_{2i}}) &= \frac{1}{(1-\rho^2)} \begin{pmatrix} \lambda_{\epsilon_{1i}} & -\sqrt{\lambda_{\epsilon_{1i}}\lambda_{\epsilon_{2i}}}\rho \\ -\sqrt{\lambda_{\epsilon_{1i}}\lambda_{\epsilon_{2i}}}\rho & \lambda_{\epsilon_{2i}} \end{pmatrix}, \\
(\lambda_{\epsilon_{si}}|a_\epsilon, b_\epsilon) &\sim \mathcal{G}a(a_\epsilon^2/b_\epsilon, a_\epsilon/b_\epsilon),
\end{aligned}
$$

where ρ is the correlation between samples, $\lambda_{\epsilon_{si}}$ is the precision of gene i in sample s, and $\mathcal{G}a(a_\epsilon^2/b_\epsilon, a_\epsilon/b_\epsilon)$ denotes a Gamma distribution with mean a_ϵ and variance b_ϵ. We use an exchangeable prior for the precisions, so that information is shared between the genes. This allows shrinkage of very small and very large variances. The priors are specified in the next subsection.

3.2 Priors

We use a vague but proper prior for the precision of the random effects λ_{γ_s}, exponential with mean 200, so that $\lambda_{\gamma_s} \sim \mathcal{G}a(1, 0.005)$. Apart from the γ's all the other effects are assumed to be random with large variance, namely $\mathrm{N}(0, 25)$. They are fixed effects but are estimated in a Bayesian way, so that uncertainty about those parameters can be captured as part of the sampling process (Lindley and Smith 1972).

For identifiability, we impose the constraints $\alpha_1 = 0$ and $\beta_1 = 0$, $\delta_{11} = \delta_{12} = \delta_{21} = 0$, $\eta_1 = 0$ and $\eta_R = 0$. We also need two constraints on the γ_{is}, such as $\sum_i \gamma_{is} = 0$ for $s = 1, 2$. However, instead of including these constraints as part of the model definition, we let the $\boldsymbol{\gamma}$'s be "free" during the sampling process, and identify the parameters afterwards from the sampled values; see Section 3.3.

We also use vague but proper priors for the error precisions, namely $a_\epsilon \sim \mathcal{U}_{[0,1000]}$ and $b_\epsilon \sim \mathcal{U}_{[0,1000]}$. The prior for the correlation between the two samples is given by $\rho \sim \mathcal{U}_{[-1,1]}$.

The prior for the degrees of freedom ν_r is uniform on the set $\{1, 2, \ldots, 10, 20, \ldots, 100\}$. A similar approach is taken by Besag and Higdon (1999). They use a uniform hyperprior on the set $\{1, 2, 4, 8, 16, 32, 64\}$ for the degrees of freedom. From a practical point of the view, the biggest difference between our approach and theirs is that we also include 3 in the set of possible values of ν_r. Our results suggest this to be useful, as there can be a noticeable difference between results for low degrees of freedom, especially 2, 3 and 4, but much smaller differences for larger values of ν_r. By using a prior that allows degrees of freedom between 1 and 100, we allow a wide range of sampling errors from the heavy tailed Cauchy ($\nu = 1$) to nearly Gaussian ($\nu = 100$).

3.3 Parameter Estimation

Realizations were generated from the posterior distribution via Markov chain Monte Carlo (MCMC) algorithms (Gelfand and Smith 1990; Brooks 1998). Univariate updating was used for all parameters except $\lambda_{\epsilon_{1i}}$ and $\lambda_{\epsilon_{2i}}$, which were updated simultaneously. We used Gibbs updates when the full conditionals had a simple form; otherwise we used slice sampling with the "stepping out" procedure (Neal 2003).

The model (1) does not allow the identification of all parameters because we do not impose any constraint on $\boldsymbol{\gamma}$. However, contrasts involving elements of $\boldsymbol{\gamma}$ are identified, and one could force all the parameters to be identified by imposing constraints such as $\sum_i \gamma_{is} = 0$ for $s = 1, 2$. For simplicity we did not take such an approach. Instead we fitted the unconstrained model and post-processed the MCMC output to identify all the parameters. By post-processing, we mean that after running the MCMC algorithm, we changed the simulated values of γ_{is} so that $\sum_r \gamma_{sr} = 0$ for $s = 1, 2$, and then recomputed the corresponding other parameters at each iteration. A similar approach has been taken to solving the label-switching problem in Bayesian inference for finite mixture models using MCMC (Stephens 2000; Celeux, Hurn, and Robert 2000).

We started the Markov chain from the least squares estimates of the parameters. We used the method of Raftery and Lewis (1992, 1996) , to determine the number of iterations, based on a short pilot run of the sampler. Software to implement this is available at `http://lib.stat.cmu.edu/S/gibbsit`, `http://lib.stat.cmu.edu/general/gibbsit`, and as part of the CODA package for MCMC diagnostics. For each dataset presented here, this suggested that a sample of no more than about 50,000 iterations with 1,000 burn-in iterations was enough to estimate standard posterior quantities. Guided by this, and leaving some margin, we used 100,000 iterations with 5,000 burn-in iterations, and stored every 10th iteration after the burn-in period.

We estimated the shift κ in advance by fitting (1) with $w_{icr} \equiv 1$ and $\lambda_{\epsilon_{is}} \equiv \lambda_{\epsilon_s}$ via MCMC, and treating κ as a parameter with vague uniform prior $\kappa \sim \mathcal{U}_{[0,10000]}$. We then estimated κ by its posterior mean.

At first sight it would seem natural to estimate κ instead by including it as a parameter in the MCMC estimation of the full model (1), but we did not do so, for the following reason. If we did so, the posterior distribution of the quantities of interest, the γ_{is}, would be averaged over different values of κ. However, when κ changes, so does the scale on which γ_{is} is measured and hence its interpretation, and so this would amount to averaging quantities denoted by the same symbol, but that actually have different interpretations. We therefore opted to

estimate κ first and then estimate the other parameters conditionally on the resulting value of κ. A similar issue arises in making inference about regression parameters when a Box-Cox (1964) transformation has been used. Box and Cox (1982) pointed out that inflating the standard errors of regression parameters to take account of uncertainty about the transformation used amounts to averaging over inferences on different scales, and so is scientifically inappropriate. They recommended first estimating the transformation parameter, and then making inference about the regression parameters conditionally on the resulting estimate. In practice, in our data sets, the posterior distribution of κ was highly concentrated, and the results would have been similar had we treated κ as a parameter of the full model (1) in the MCMC estimation.

We estimate the effect of gene i in sample s by the posterior mean of γ_{is}, namely

$$\bar{\gamma}_{is} = \frac{1}{K}\sum_{k=1}^{K} \gamma_{is}^{(k)},$$

where K is the number of iterations used in the MCMC after burn-in. Interest often focuses on the difference between the two samples, i.e. on the log ratio of the intensities, which can be estimated by $(\bar{\gamma}_{i1} - \bar{\gamma}_{i2})$ for gene i. The simplest estimate of the ratio of the log intensities is the raw log ratio, namely $(\overline{z}_{i1cr} - \overline{z}_{i2cr})$, where $z_{iscr} = \log_2(y_{iscr} - x_{iscr})$ denotes the raw log measurements. Another popular estimate is the ANOVA normalized log ratio, $(\hat{\gamma}_{i1} - \hat{\gamma}_{i2})$, where $\hat{\gamma}_{is}$ is the least squares estimate of the effect of gene i in sample s from our general model applied to the raw log measurements. The name "ANOVA normalization" was introduced by Kerr, Martin, and Churchill (2000), and we will use the same terminology even though our model is slightly different. Another much used estimate is the lowess normalized log ratio computed by carrying out a nonlinear normalization of each individual array (Dudoit, Yang, Callow, and Speed 2002). The log ratios of a given array are transformed in the following way:

$$z_{i1cr} - z_{i2cr} \Leftarrow z_{i1cr} - z_{i2cr} - h_r(z_{i1cr} + z_{i2cr}),$$

where $h_r(.)$ is the lowess fit to the plot of the log ratios $(z_{i1cr} - z_{i2cr})$ against the overall intensity $(z_{i1cr} + z_{i2cr})/2$ (Yang, Dudoit, Luu, Lin, Peng, Ngai, and Speed 2002). Then the normalized log ratios are averaged across replicates. The different estimates are compared in Section 4.2.

Another estimation approach, applicable if three or more replicates are available, is to apply a statistical test to the normalized measurements so as to remove outliers. Ideker,

Thorsson, Siegel, and Hood (2000) filter the normalized measurements to remove outliers by Dixon's test (Kanji 1993), with significance level $\alpha = 0.1$. For each individual gene, the measurements are filtered in each sample independently. In the next section, we compare Dixon's test to our hierarchical t-formulation on 3 different genes. In Section 4.2, we assess the effect of Dixon's test on the variability between estimates from replicated experiments.

4 RESULTS

4.1 Illustrative Results

In this section, we use specific genes from the three experimental datasets described in Section 2 to illustrate our model. The algorithm presented in Section 3.3 was used to fit the general model to the like and like data and the two HIV data sets. Table 1 summarizes the estimated coefficients when the model is applied to the HIV1 data. The posterior modes of the degrees of freedom of the t-distribution, ν_r, ranged from 5 to 10, indicating that the sampling errors are heavier-tailed than the Gaussian distribution. There is substantial between-sample correlation, estimated as 0.68, even after removing design effects and gene effects. Our model also captures the nonconstant variance with posterior means 7.10 and 1.75 for a_ϵ and b_ϵ respectively. The posterior mean of 1.75 for the variance of the gene precisions, b_ϵ, allows the gene precisions to be quite different and should capture the larger variance at low intensity.

Tables 2–4 illustrate the effect of the t-distribution when outliers are present. The weights correspond to the posterior mean of the w_{icr} for each pair of observations. Conditioning on the w_{icr}'s, the posterior mean can be seen as a weighted mean with the w_{icr} as weights. Table 2 shows that Dixon's test as applied by Ideker, Thorsson, Siegel, and Hood (2000), i.e. on the raw measurements in each sample independently, fails to remove a clear outlier. Because of the outlier, the difference between the estimates of the effect of the same gene from the HIV1 data and the HIV2 data is quite large, a difference of -0.71. Our model clearly downweights the outlier, and as a result the difference between the two estimates is much smaller with our method, at 0.12. The median also performs well in this case.

In Table 3, there is a clear outlier in the HIV1 data that is removed by Dixon's test and is also downweighted by our model. However the HIV2 experiment also seems to contain an outlier that is not removed by Dixon's test. In this case our method gives estimates that differ between experiments by 0.41, while the other three methods yield estimates that differ by more than three times as much.

Table 1: Summary of the Coefficients from our Baysian Model on the HIV1 Data.

Parameter	Effect	Bayesian estimate	Posterior sd	$q_{0.025}$	$q_{0.975}$
μ	baseline intensity	7.98	0.006	7.97	7.99
α_2	sample effect	-0.28	0.003	-0.29	-0.28
β_2	dye effect	-0.14	0.009	-0.16	-0.12
δ_{22}	dye×sample interaction	0.58	0.005	0.57	0.59
η_2	array 2 effect	-0.47	0.008	-0.48	-0.45
η_3	array 3 effect	-0.06	0.009	-0.08	-0.04
λ_{γ_1}	gene precision sample 1	0.34	0.007	0.33	0.35
λ_{γ_2}	gene precision sample 2	0.34	0.007	0.33	0.35
ρ	correlation between samples	0.68	0.005	0.67	0.69
a_ϵ	mean of error precisions	7.10	0.12	6.89	7.32
b_ϵ	variance of error precisions	1.75	0.62	0.62	3.02
ν_1	df for array 1	10	0.5	9	10
ν_2	df for array 2	9	0.6	8	10
ν_3	df for array 3	6	0.3	5	6
ν_4	df for array 4	5	0.1	4	5

Note: The Bayesian estimate is the posterior mean, except for ν_1, ν_2, ν_3 and ν_4, for which it is the posterior mode.
$q_{0.025}$: 0.025 quantile. $q_{0.975}$: 0.975 quantile

Table 2: Log Ratios of One Gene of the HIV Data Sets. Dixon's test (Ideker, Thorsson, Siegel, and Hood 2000) fails to remove a clear outlier.

	Replicates						posterior	The Dixon
	1	2	3	4	mean	median	mean	mean
HIV 1 (log ratio)	-3.21	-0.04	0.34	0.41	-0.63	0.15	0.14	-0.63
weights	0.06	0.92	1.06	1.07				
HIV 2 (log ratio)	-0.20	0.08	0.15	0.31	0.08	0.11	0.02	0.08
weights	1.01	1.14	1.15	1.03				
difference					-0.71	0.04	0.12	-0.71

Note: ⋆gene removed by Dixon's test on the normalized measurement at the 10% level.
The Dixon mean corresponds to the sample mean after removing the outlier (if any).

Table 3: Log Ratios of One Gene of the HIV Data Sets. Dixon's test removes a clear outlier.

	Replicates						posterior	The Dixon
	1	2	3	4	mean	median	mean	mean
HIV 1 (log ratio)	-0.94	0.30	0.75	4.61*	1.08	0.53	-0.07	-0.11
weights	0.46	1.06	1.00	0.05				
HIV 2 (log ratio)	-4.04	-2.51	-0.16	0.08	-1.65	-1.33	-0.48	-1.65
weights	0.52	0.79	1.04	0.99				
difference					2.73	1.86	0.41	1.54

Note: *gene removed by Dixon's test on the normalized measurement at the 10% level. The Dixon mean corresponds to the sample mean after removing the outlier (if any).

In Table 4, Dixon's test incorrectly removes non-outlying replicates. The effect on the variability of the estimates is large. Our method smooths the outliers and once again reduces the variability between estimates.

Table 4: Log Ratios of One Gene of the HIV Data Sets. Non-outlying replicates are incorrectly removed by Dixon's test.

	Replicates						posterior	The Dixon
	1	2	3	4	mean	median	mean	mean
HIV 1 (log ratio)	0.08	0.21	0.52*	0.87*	0.42	0.36	0.28	0.14
weights	1.20	1.16	1.14	0.93				
HIV 2 (log ratio)	-1.44	0.68	1.63	1.95	0.71	1.16	0.48	0.71
weights	0.57	0.99	0.99	0.89				
difference					-0.29	-0.80	-0.20	-0.57

Note: *gene removed by Dixon's test on the normalized measurement at the 10% level. The Dixon mean corresponds to the sample mean after removing the outlier (if any).

4.2 Between-Replicate Variability of Estimates

In this section, we compare the different log ratio estimates introduced in Section 3.3. We first compare our estimates with the ANOVA normalized log-ratios for the like and like data, as shown in Figure 5. These are a natural first point of comparison, because they are essentially the raw log ratios, normalized and corrected for design effects. In theory, the like and like data should not show any differentially expressed genes. The gray lines in Figure 5

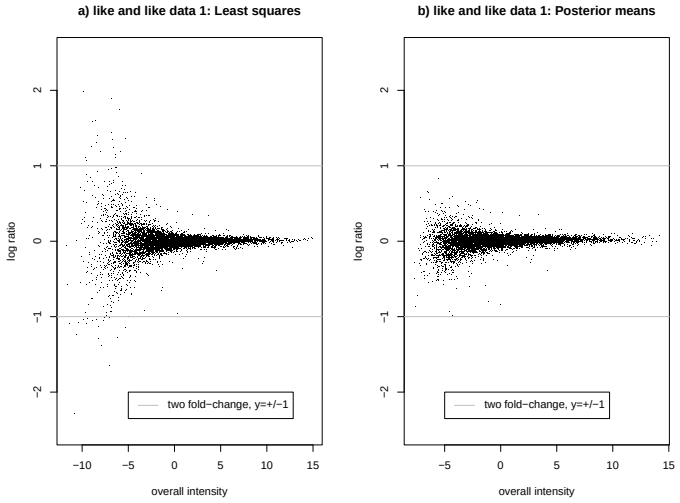

Figure 5: Log Ratio Estimates as a Function of the Overall Intensity on the Like and Like Data (first 4 replicates). The gray lines show a two-fold change. The number of false positives for the normalized log ratios is 36, as against only 0 for the posterior means, a 100% reduction.

show a two-fold change. The ratio of two is sometimes used as a rule of thumb for selecting differentially expressed genes (Yang, Ross, Kuang, Brown, and Weigel 1999). Because of the high variability at low intensity, some of the genes show a greater than two-fold change in expression. Using the two-fold change rule, the number of false positives for the ANOVA normalized log ratios is 36, as against 0 for the posterior means, an 100% reduction.

We highlighted two groups of genes in the two HIV data sets. The first group consists of HIV genes (positive controls) that are known to be differentially expressed, and the second one consists of non-human genes (negative controls), which are known not to be differentially expressed. Figure 6 shows that our model enhances the identification of the differentially expressed genes. It shrinks the low intensity (highly variable) genes, but does not modify the differentially expressed genes too much.

We now compare all the methods described in Section 3.3 by dividing the like and like data and the HIV data into two groups of four replicates, computing an estimate of the log ratio for each gene from each of the two groups, and computing the sum of squared differences between the two estimates, summed over all genes. The eight replicates of the HIV data were separated into two groups of four, consisting of the HIV1 and HIV2 replicates respectively. Even though those two groups are biological rather than technical replicates, we expect the log ratios from each group to be similar.

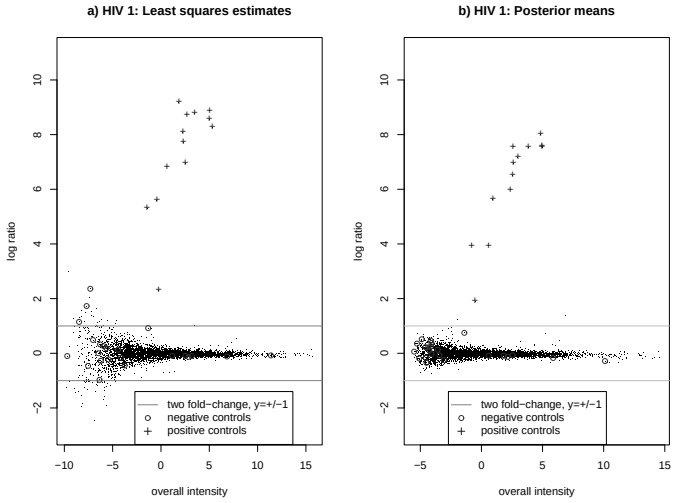

Figure 6: Log Ratio Estimates as a Function of the Overall Intensity on the HIV1 Data. The gray lines show a two-fold change. The number of false positives at low intensity is greatly reduced. The log ratio estimates of the true differentially expressed genes stay about the same.

The raw log ratios performed poorly, with high variability between estimates. The ANOVA normalized log ratios performed better, as shown in Tables 5 and 6. Removing outliers using Dixon's test as in Ideker, Thorsson, Siegel, and Hood (2000) did not improve things; it actually did worse than the normalized log ratios that take no account of outliers.

The between-replicate variability is greatly decreased when using our model. Our method reduced the sum of squared differences between estimates by 64% for the like and like data and by 83% for the HIV data, compared to the raw log ratios. Our method also performed considerably better than the other methods considered, by this measure. For the like and like data, it produced a sum of squared differences between estimates that was 55% less than that for the best competing method (in that case the lowess normalized log ratio). For the HIV data, it was 58% better than the best competing method (in that case the ANOVA normalized log ratio).

We also compared our approach to the quality filtering of Tseng, Oh, Rohlin, Liao, and Wong (2001). They filter so-called low quality genes, based on their coefficient of variation. A gene whose coefficient of variation is too large is removed from the dataset. When we used their software on our data, it removed about 1000 genes in each of the two groups of the HIV data. The software can be downloaded at `http://biosun1.harvard.edu/~tseng/download.html`. The numbers of genes removed in each group were different and therefore it was not possible

Table 5: Sum of Squared Differences Between the Estimates when Dividing the Like and Like Data in Two Groups of 4 Replicates. The posterior mean reduces the variability by 64% relative to the raw log ratios.

Estimates	SSD	$SSD/(SSD$ raw log ratio)
Raw log ratio	473	1.00
ANOVA normalized log ratio	430	0.91
ANOVA normalized log ratio (w/ Dixon)	560	1.18
Lowess normalized log ratio	381	0.81
Posterior mean	170	0.36

Note: ANOVA normalized log ratios are computed by subtracting the least squares estimates of the design effects. Lowess normalized log ratio are computed by doing a Lowess normalization to each individual array. ANOVA normalized log ratio (w/ Dixon): Dixon's test at the 10% level was used to remove outliers.

Table 6: Sum of Squared Differences Between the Estimates when Dividing the HIV Data in Two Groups of 4 Replicates. The posterior mean reduces the variability by 83% relative to the raw log ratios.

Estimates	SSD	$SSD/(SSD$ raw log ratio)
Raw log ratio	1941	1.00
ANOVA normalized log ratio	778	0.40
ANOVA normalized log ratio (w/ Dixon)	871	0.44
Lowess normalized log ratio	814	0.42
Posterior mean	326	0.17

Note: Least squares normalized log ratios are computed by subtracting the least squares estimates of the design effects. ANOVA normalized log ratio are computed by doing a Lowess normalization to each individual array. ANOVA normalized log ratio (Dixon): Dixon's test at the 10% level was used to remove outliers.

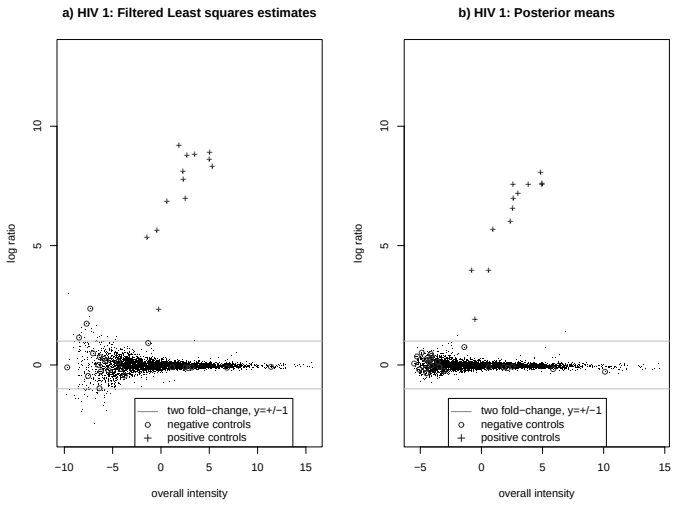

Figure 7: Log Ratio Estimates as a Function of the Overall Intensity on the HIV1 Data. The gray lines show a two-fold change. (a) Normalized log ratios after filtering genes as in Tseng, Oh, Rohlin, Liao, and Wong (2001). (b) Log ratio estimates from our model. The filtering did not remove the poor quality genes at low intensity.

to compare their method with ours directly in terms of variability between estimates. Figure 7 compares the estimated log ratios from our model to the least squares normalized log ratios after filtering. It shows that the filtering did not necessarily remove the low intensity genes. In their paper, Tseng et al. (2001) also recommend looking for replicate outliers when a spot fails to pass the quality filtering. However they do not mention how to decide if a replicate is an outlier. This is a crucial point especially when the number of replicates is small.

4.3 Should the Background be Subtracted?

The image from which individual gene expression levels are estimated takes the form of roughly circular spots superimposed on a background. Expression levels are estimated by measuring the average intensity in the spot. The measured intensity in the background is often greater than zero for various physical reasons, including flourescence of the glass substrate, amplifier offset, dark current, and so on. Thus the estimate of the intensity in a spot is often modified by subtracting the estimated background intensity. In general, background intensities vary spatially within a slide, and so it is common to estimate the background for each spot separately (e.g. Yang et al. 2000). Brown, Goodwin, and Sorger (2001) have argued that local background is a poor measure of nonspecific flourescence, and

that one should use a global background estimate, i.e. use the same background estimate for all the genes on a slide.

However, some authors have pointed out that subtracting the background can have the negative effect of increasing the variability, especially at low intensity (Rocke and Durbin 2001; Yang, Buckley, Dudoit, and Speed 2002; Cui, Kerr, and Churchill 2002; Glasbey and Ghazal 2003). A counterargument to this is that by not subtracting the background one increases all the intensity measurements, and so one tends to reduce the estimates of ratios that are large, thus biasing the ratio estimates of differentially expressed genes downwards. This debate about whether or not to subtract the background remains unresolved.

A comparison of standard (ANOVA normalized) log ratios with and without background subtraction (Figures 6 and 8) shows that the arguments on both sides of the debate are correct for our data. Figure 6(a) shows the estimates with background subtraction, and the high variance at low intensities is clear; using the two-fold change rule of thumb, this seems likely to lead to a considerable number of false positive assessments of differential expression. Figure 8 shows the same plot, but without background subtraction. The variance is indeed considerably reduced. But this comes at a high price in terms of bias. For the 13 genes known to be differentially expressed, the median log ratio is 8.1 with background subtraction, and 4.4 without. Such a level of bias could lead us to miss genes that are moderately differentially expressed, and indeed one of our differentially expressed genes falls below the two-fold threshold when the background is not subtracted. On the other hand, three of the 29 genes known not to be differentially expressed exceed the threshold when the background is subtracted, but none do so when the background is not subtracted.

Inspection of the posterior means in Figure 8 suggests that our method allows one to have the best of both worlds: one can subtract the background without paying such a high price in terms of variance. After background subtraction, the variance at low intensities is much less than with the standard nonrobust method, and the median log ratio for the 13 differentially expressed genes is 6.9, much closer to the nonrobust estimates with background subtraction than without. With our method, all the 13 known differentially expressed genes exceed the two-fold threshold, while none of the 29 genes known not to be differentially expressed do.

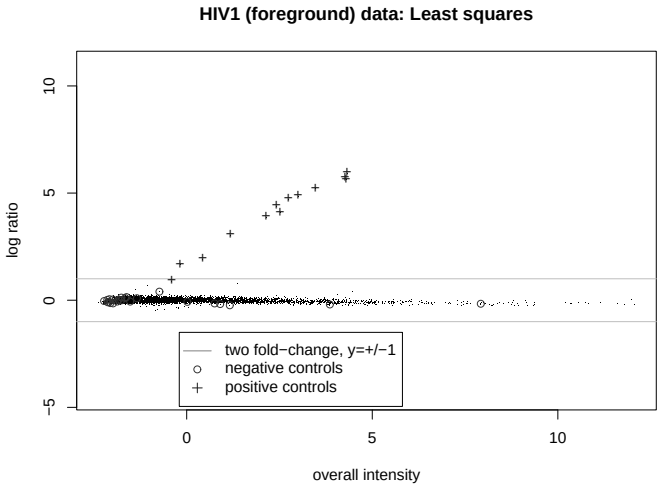

Figure 8: ANOVA Normalized Log Ratio Estimates Without Background Subraction for the HIV Data. Not subtracting the background shrinks the estimates towards the x-axis. One of the genes known to be differentially expressed shows a less than two-fold change.

5 DISCUSSION

We have developed a Bayesian hierarchical model for estimating cDNA microarray intensities in a way that is robust to outlying measurements caused by things such as scratches, dust, imperfections in the glass and imperfections in the array production. The robustness is achieved by using a hierarchical t-distribution and allowing the data to choose the number of degrees of freedom. Our model borrows strength from all the genes when deciding if a gene is an outlier. This is essential: it is hard to detect outliers based only on the four measurements for a single gene. Classical robust estimators, such as M-estimators, would be inefficient with a small number of replicates. For example a trimmed mean with 4 replicates would remove at least 2 observations and the estimate would then be based on 2 replicates. Our model works well with four replicates, thanks to the borrowing of strength. Our model also deals with the classical issues of design effects, normalization, transformation and nonconstant variance.

We decided to model the log transformed microarray intensities. We have specified our model on the scale of log transformed intensities. Durbin, Hardin, Hawkins, and Rocke (2002) and Huber, von Heydebreck, Sultmann, Poustka, and Vingron (2002) independently proposed a transformation that stabilizes the variance very well. However, this transformation is somewhat complex, and Rocke and Durbin (2003) have shown that the shifted log

transformation provides a good approximation while keeping the ease of interpretation of log ratios.

We estimate the dye bias assuming that most of the genes are not differentially expressed. This assumption is usually at least approximately correct, and without it one cannot distinguish a poor RNA preparation from differential expression. If one does not accept this assumption, the term α, the dye effect, should be removed from the model. If we still want to estimate the sample bias, a technique similar to that of Li and Wong (2000) where a group of genes that are believed not to be differentially expressed is selected, could be used. Our model assumes that the data are the results of a dye swap experiment. If no dye swap was performed, the sample effect and the dye effect will be confounded, and the dye effect should be removed from the model. Using a dye swap experiment, we can estimate the dye effect and normalize the data in a linear fashion. The difference in the dye effects tends to depend on the gene and the intensity, resulting in a nonlinear effect, and this can be removed by normalizing the data before hand using the lowess normalization (Yang, Buckley, Dudoit, and Speed 2002).

In our application we have considered only balanced dye-swap experiments. However, our model could easily be modified to take account of other designs such as those proposed by Kerr and Churchill (2001) and Dobbin, Shih, and Simon (2003). For example, our model could be extended to the loop design introduced by Kerr and Churchill (2001). Note that a loop design with only two samples (or varieties) is just a dye swap experiment, and so a more general loop design is a natural extension of the present work.

References

Besag, J. E. and D. M. Higdon (1999). Bayesian analysis of agricultural field experiments (with discussion). *Journal of the Royal Statistical Society, Series B 61*, 691–746.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B 26*, 211–252.

Box, G. E. P. and D. R. Cox (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association 77*, 209–210.

Brooks, S. P. (1998). Markov chain Monte Carlo and its application. *The Statistician 47*, 69–100.

Brown, C. S., P. C. Goodwin, and P. K. Sorger (2001). Image metrics in the statistical

analysis of DNA microarray data. *Proceedings of the National Academy of Science 98*, 8944–8949.

Celeux, G., M. Hurn, and C. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association 95*, 957–970.

Chen, Y., E. R. Dougherty, and M. L. Bittner (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics 2*, 364–374.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of American Statistical Assiciation 74*, 829–836.

Cui, X., M. K. Kerr, and G. A. Churchill (2002). Data transformations for cDNA microarray data. Technical report, The Jackson Laboratory.

Dobbin, K., J. H. Shih, and R. Simon (2003). Questions and answers for indentifying differentially expressed genes. *Journal of the National Cancer Institute 95*, 1362–1369.

Dudoit, S., Y. H. Yang, M. J. Callow, and T. P. Speed (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica 12*, 111–139.

Durbin, B., J. Hardin, D. Hawkins, and D. M. Rocke (2002). A variance-stabilizing transformation for the gene-expression microarray data. *Bioinformatics 18*, 105S–110S.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*, 398–409.

Glasbey, C. A. and P. Ghazal (2003). Combinatorial image analysis of DNA microarray features. *Bioinformatics 19*, 194–203.

Gottardo, R., J. A. Pannucci, C. R. Kuske, and T. Brettin (2003). Statistical analysis of microarray data: A Bayesian approach. *Biostatistics 4*, 597–620.

Huber, W., A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics 18*, S96–S104.

Ideker, T., V. Thorsson, A. F. Siegel, and L. E. Hood (2000). Testing for differentially expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology 7*, 805–817.

Kanji, G. K. (1993). *100 Statistical Tests*. Sage.

Kerr, M. K. and G. A. Churchill (2001). Experimental design for gene expression microarrays. *Biostatistics 2*, 183–201.

Kerr, M. K., M. Martin, and G. Churchill (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology 7*, 819–837.

Lewin, A., S. Richardson, C. Marshall, A. Glazier, and T. Aitman (2003). Bayesian modelling of differential gene expression. Technical report, Imperial College, London.

Li, C. and W. H. Wong (2000). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science 98*, 31–36.

Lindley, D. V. and A. F. M. Smith (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B 34*, 1–41.

Lönnstedt, I. and T. P. Speed (2002). Replicated microarray data. *Statistica Sinica 12*, 31–46.

Neal, R. (2003). Slice sampling. *The Annals of Statistics 31*, 705–767.

Newton, M. C., C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology 8*, 37–52.

Raftery, A. E. and S. M. Lewis (1992). How many iterations in the Gibbs sampler? In *Bayesian Statistics 4* (edited by J. M. Bernardo et al.), pp. 763–773. Oxford: Oxford University Press.

Raftery, A. E. and S. M. Lewis (1996). Implementing MCMC. In *Markov Chain Monte Carlo in Practice.* (edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter), Chapter 7, pp. 115–130. London: Chapman and Hall.

Rocke, D. M. and B. Durbin (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology 8*, 557–569.

Rocke, D. M. and B. Durbin (2003). Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics 19*, 966–972.

Stephens, M. (2000). Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society, Series B 62*, 795–809.

Tseng, G. C., M. Oh, L. Rohlin, J. C. Liao, and W. H. Wong (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assesement of gene effects. *Nucleic Acid Research 29*, 2549–2557.

Tukey, J. W. (1957). On the comparative anatomy of transformations. *Annals of Mathematical Statistics 28*, 602–632.

van't Wout, A. B., G. K. Lehrma, S. A. Mikheeva, G. C. O'Keeffe, M. G. Katze, R. E. Bumgarner, G. K. Geiss, and J. I. Mullins (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of $CD4^+ - T -$ Cell lines. *Journal of Virology 77*, 1392–1402.

Yang, G. P., D. T. Ross, W. W. Kuang, P. O. Brown, and R. J. Weigel (1999). Combining SSH and cDNA microarrays for rapid identification of differentially expressed genes. *Nucleic Acid Research 27*, 1517–1523.

Yang, Y. H., M. J. Buckley, S. Dudoit, and T. P. Speed (2002). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics 11*(1), 108–136.

Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed (2002). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research 30*, e15.

Yeung, K. Y., C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics 17*, 977–987.